The Annals of Applied Statistics
2020, Vol. 14, No. 4, 2037–2052
https://doi.org/10.1214/20-AOAS1379

MODELLING THE SOUND PRODUCTION OF NARWHALS USING A POINT PROCESS FRAMEWORK WITH MEMORY EFFECTS

BY ALEKSANDER SØLTOFT-JENSEN[1,*], MADS PETER HEIDE-JØRGENSEN[2] AND SUSANNE DITLEVSEN[1,†]

[1]*Data Science Laboratory, Department of Mathematical Sciences, University of Copenhagen,* [*]*soltoft-jensen@math.ku.dk;* [†]*susanne@math.ku.dk*

[2]*Department of Birds and Mammals, Greenland Institute of Natural Resources, mhj@ghsdk.dk*

Obtaining an adequate description of the behaviour of narwhals in a pristine environment is important to understand natural behaviour as well as providing the means to determine potential changes in behaviour directly or indirectly caused by human activity. Based on Acousonde$^{\text{TM}}$ data from five narwhals in Scoresby Sound, this paper aims at modelling buzzing and calling rates of East Greenland narwhals as functions of time, space and, possibly, autoregressive memory. Both buzzing and calling are sounds produced by narwhals. Buzzing is a way for the whale to navigate and locate prey using echolocation, while calling is associated with social communication between whales. Logistic regression models without and with autoregressive components are compared based on AIC and comparatively assessed using diagnostics from point process theory. Adding an autoregressive component appears to improve the models, and further improvements for the buzzing model are made with a non-GLM extension. Effects of extrinsic covariates and memory are presented and interpreted. Buzzing occurs at deeper depths, and initiations of buzzes are separated by refractory periods. A possible feeding area is identified. Calling occurs closer to the surface, and, while the probability of calling in general is lower than buzzing, it is more likely that calls are clustered together rather than spread randomly.

1. Introduction. The narwhal (*Monodon monoceros*) is one of the deepest diving cetaceans with the maximum exceeding 1800 m (Heide-Jørgensen et al. (2015)), with the largest abundances found in East and West Greenland and in the Canadian High Arctic (Heide-Jørgensen et al. (2002)). The narwhals dive to forage and depend on acoustics for sensing their environment, navigating and capturing prey at depth (Rasmussen, Koblitz and Laidre (2015)). Sound plays a crucial role in the life of East Greenland narwhals, because they dive to depths much below the photic zone and are seasonally exposed to darkness or limited daylight as well as extensive ice coverage (Blackwell et al. (2018)). To navigate and feed, a narwhal can buzz which is a form of echolocation. Calling, on the other hand, is used for communication with conspecifics. Establishing adequate models that describe buzzing and calling behaviour is thus a significant step toward attaining knowledge of how narwhals behave. The data are obtained from narwhals in a relatively pristine environment, and one could, therefore, hope that resulting models are representative of the whales' normal behaviour.

Climate changes decrease the sea ice coverage, exposing the natural habitat of narwhals to anthropogenic factors like underwater noise from shipping and seismic exploration (Koblitz et al. (2016)). As noise pollution is predicted to increase, the models developed here have potential to be used as a reference when trying to determine if noise from human activities

Received October 2019; revised May 2020.
Key words and phrases. Point process, narwhal, buzz and call, logistic regression with memory, autoregressive process, behavioural data of marine mammals, ecology.

alters the whales' behaviour. We will qualitatively and quantitatively describe the vocal activities to understand the behaviour of narwhals under natural conditions and to ensure the long-term conservation of one of the most specialized species in the North Atlantic.

In most acoustic studies of narwhals, data were collected with dipping hydrophones, autonomous passive acoustic recorders or hydrophone arrays. Such studies are generally limited, with little information on the spatial and temporal variation in sound production of specific individuals, since they are stationary. The development of animal-borne acoustic recorders has opened up new possibilities for monitoring and gaining understanding of individual acoustic behaviour of freely moving cetaceans (Johnson, Aguilar de Soto and Madsen (2009)). Moreover, only recently it has become possible to obtain attachment durations of various days or weeks which provides rich information about the variability over time for whales that are frequenting different habitats. Therefore, it is important to develop robust statistical methods to deal with these new types of large data sets.

In Blackwell et al. (2018), the acoustic behaviour of East Greenland narwhals was described extensively, possibly for the first time. Various descriptive statistics were presented to gain insight into the spatial and temporal patterns of sound production, where clicks, buzzes and calls—all being sounds produced by narwhals—were considered. Additionally, logistic regression models were suggested for five out of six available whales. In these models, conditional probabilities of buzzing and calling were estimated given three extrinsic predictors: *area*, *depth* and *time of day*; the first being categorical, and the latter two being continuous. An *area*-effect as well as nonlinear effects of *depth* and *time of day* were justified based on statistical significance. The regression rested on the assumption that observations each second were independent, given the extrinsic predictors. Using the same data set, this paper presents two models that relax this assumption. The best model is selected through AIC and compared with the previously suggested logistic regression model (hereafter denoted the base model) with respect to diagnostics in a point process framework which previously has been used in neuroscience; see Truccolo et al. (2005), Kass, Eden and Brown (2014, Chapter 19), and Li et al. (2016). To see how the effects of the extrinsic covariates and, therefore, the interpretations possibly change under the different assumptions, the base model and best model are presented together for each whale.

Behavioural data obtained from tagged cetaceans like the depth measurements and GPS positions used as covariates here have typically been analyzed with hidden Markov models (DeRuiter et al. (2017), Langrock et al. (2014), Ngo, Heide-Jørgensen and Ditlevsen (2019), Quick et al. (2017)); see also Patterson et al. (2017) for a review on statistical methods used for this type of data. Spatial point process models have been used to model large-scale survey data of cetaceans (Yuan et al. (2017)); however, these data are not from animal-borne tags and can thus not model detailed individual behaviour. The point process approach taken here to model acoustic behaviour of cetaceans is new and provides a flexible modelling framework that incorporates individual memory components and covariates.

2. Data. Five narwhals were live-captured and instrumented with Acousonde (www.acousonde.com) acoustic recorders in Scoresby Sound, East Greenland, in August 2013–2016 (Figure 1, see Blackwell et al. (2018) for details on capturing and tagging). The Acousonde recorders were released from the whales after three to eight days of attachment and were retrieved at sea by radio tracking. The Acousonde recorders provided archived data on depth for every 1 s, continuous recordings of whale vocalizations and daily positions of the whales (Figures 1–2). The recordings were analyzed for presence of buzzes and calls by a combination of manual and automatic detections (Blackwell et al. (2018)). The whales showed a period of posttagging silence that was omitted from the analyses.

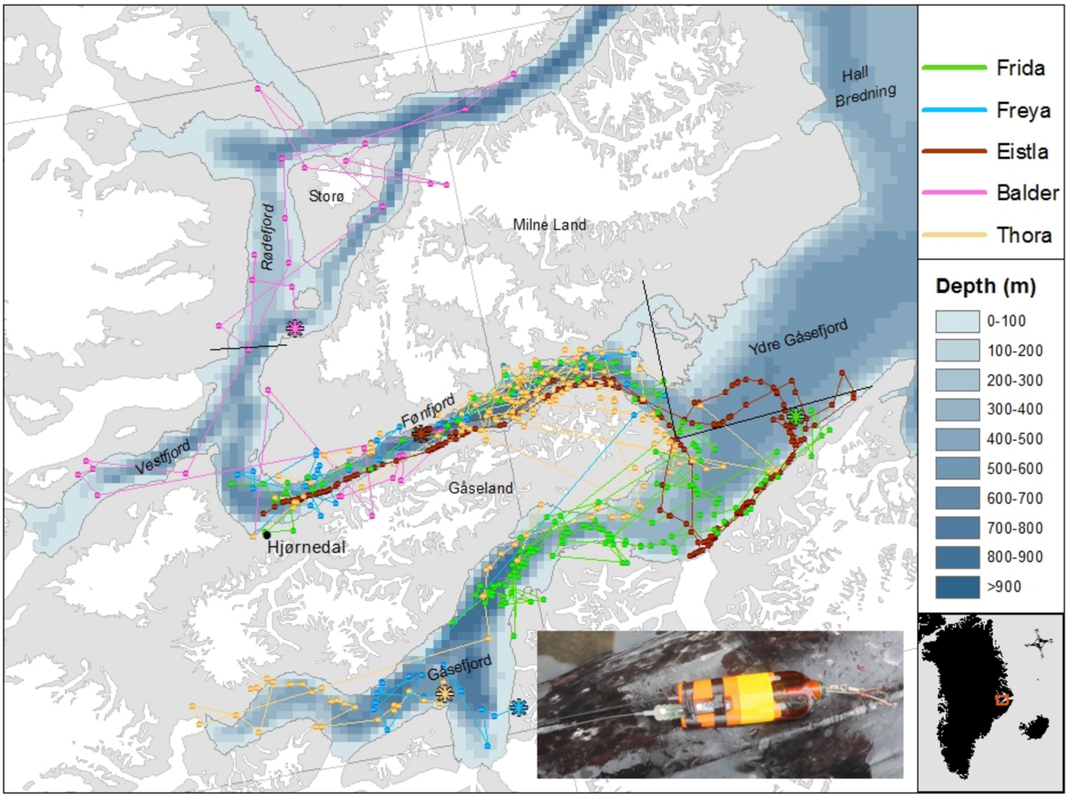

FIG. 1. *Satellite tracks of the five whales and an image of an Acousonde on a narwhal. All tracks begin near Hjørnedal and end at the star symbols.*

3. Methods. The point process framework, for example as used in Truccolo et al. (2005) and Kass, Eden and Brown (2014, Chapter 19), is chosen for this article. Here, it is assumed that the true data generating mechanism for the sounds produced by the narwhals is a point process which is a set of discrete events occurring in continuous time. For the purpose of this article, these events (henceforth referred to as event times) are the beginnings of a buzz or a call. Given an observation interval $(0, T]$, the event times are denoted $u_1, \dots, u_J$, satisfying the inequalities $0 < u_1 < \cdots < u_J \leq T$.

For $t \in (0, T]$, a point process can be completely characterized by its conditional intensity function,

$$\lambda\bigl(t \mid X(t)\bigr) = \lim_{\Delta \to 0} \frac{P(N(t+\Delta) - N(t) = 1 \mid X(t))}{\Delta}, \tag{3.1}$$

where $N(t)$ is the number of events in the interval $(0, t]$ and $X(t)$ contains information on all past events and all covariate values in $(0, t)$. Thus, for Δ small enough it follows that

$$\lambda\bigl(t \mid X(t)\bigr)\Delta \approx P\bigl(N(t+\Delta) - N(t) = 1 \mid X(t)\bigr). \tag{3.2}$$

This means that $\lambda(t \mid X(t))\Delta$ is an approximation of the conditional probability of observing an event in the interval $(t, t + \Delta]$.

The available data are based on a 1 Hz sampling rate, $\Delta = 1$ sec, and the observations are mathematically formalized as $(Y_1, X_1), \dots, (Y_T, X_T)$, where X_t is a vector of covariates and Y_t is a binary response variable with support $\{0, 1\}$, indicating whether a sound was initiated in the interval $(t, t + \Delta]$. The realization $Y_t = 1$ corresponds to the beginning of a buzz or a call at time t. Theoretically, the data can be interpreted as a discretized representation of the

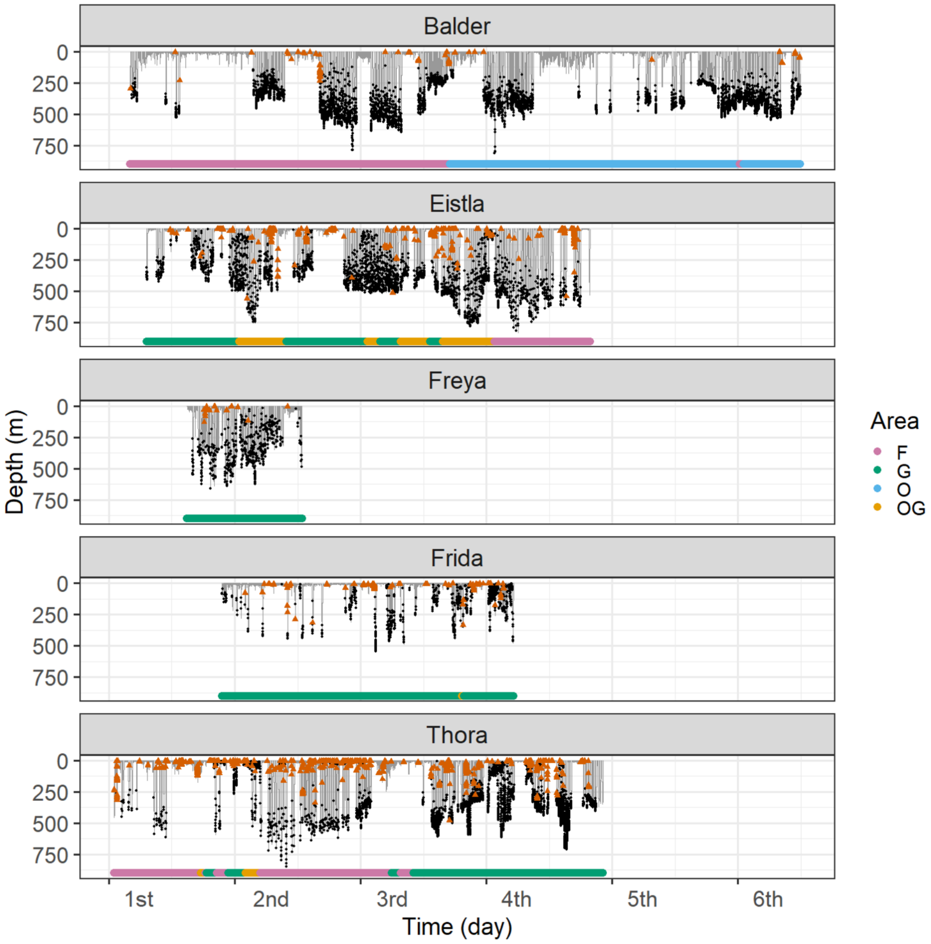

FIG. 2. *Dive depth as a function of time in days (grey lines) for each whale. The first day is defined as the day where the record starts, and ticks are placed at midnight. Buzzes are indicated with black dots and calls with triangles. Area is indicated by the bars below the depth records (colour online) (F: Fønfjord, G: Gåsefjord, O: Øfjord, OG: Outer Gåsefjord).*

true event times, and the analysis relies on assumption (3.2), implying that Δ is so small that $P(N(t+\Delta) - N(t) > 1 \mid X(t)) \approx 0$.

As mentioned in the Data section, the whales showed a period of posttagging silence, which was excluded from the analysis, such that $t = 0$ corresponds to the time point where the analysis started. Therefore, for some Q being smaller than this period, it is always the case that $y_{-Q+1} = y_{-Q+2} = \cdots = y_0 = 0$, where y_t is the observation of Y_t; this information is included and conditioned on in the model. Implicitly conditioning on all other relevant predictors and using p as a generic symbol for a probability mass function (pmf), the pmf of the data can be written as

$$p(y_1, \dots, y_T \mid y_{-Q+1}, \dots, y_0) = \prod_{t=1}^{T} p(y_t \mid y_{-Q+1}, \dots, y_{t-1}). \tag{3.3}$$

As will become clear when the models are specified, it is further assumed that, for $t \in \{1, \dots, T\}$,

$$p(y_t \mid y_{-Q+1}, \dots, y_{t-1}) = p(y_t \mid y_{t-Q}, \dots, y_{t-1}), \tag{3.4}$$

implying that all relevant information of past events is contained in the preceding time period of length Q. It follows that the pmf of the data is

$$p(y_1, \dots, y_T \mid y_{-Q+1}, \dots, y_0) = \prod_{t=1}^{T} p(y_t \mid y_{t-Q}, \dots, y_{t-1}). \tag{3.5}$$

Moreover, as the responses are binary,

$$p(y_1, \dots, y_T \mid y_{-Q+1}, \dots, y_0) = \prod_{t=1}^{T} p_t^{y_t}(1 - p_t)^{1-y_t}, \tag{3.6}$$

where $p_t = P(Y_t = 1 \mid X_t) = P(N(t + \Delta) - N(t) \geq 1 \mid X(t))$. From this, the likelihood can be established through the model for p_t.

For each whale separately, three nested models of p_t are considered: M_0, M_1 and M_2, such that $M_0 \subseteq M_1 \subseteq M_2$. The base model M_0 is characterized as follows:

$$M_0 : \operatorname{logit}(p_t) = \eta_t, \tag{3.7}$$

$$\eta_t = g_A(a_t) + g_D(d_t) + g_H(h_t). \tag{3.8}$$

Above, g_A, g_D and g_H are parametric functions of the predictors $a_t \in \{F, G, O, OG\}$, $d_t \in \mathbb{R}_{\geq 0}$ and $h_t \in [0, 24)$, representing the area, depth and time of day. The intensity in this model is independent of past history, and M_0 is an inhomogeneous Poisson process. Let $\beta = (\beta_1, \dots, \beta_p)^T$ denote the parameter vector of the regression coefficients in η_t in (3.8), and let $\theta_0 = \beta$ denote the parameter vector to be estimated in M_0.

To help the reader better understand the intuition behind M_1, the largest model M_2 is introduced first. Here, an autoregressive component of order Q is added to the predictor η_t,

$$M_2 : \operatorname{logit}(p_t) = \eta_t + \sum_{q=1}^{Q} c_q y_{t-q}, \tag{3.9}$$

where $y_{t-1}, \dots, y_{t-Q}$ are past realizations of the response variables and $c = (c_1, \dots, c_Q)^T$ is the vector of autoregressive coefficients. In M_2, $\theta_2 = (\beta^T, c^T)^T$. Given a certain response (buzzing or calling), the autoregressive order Q is chosen based on the pooled AIC-value $\sum_{i=1}^{5} -2 \log L_{i,Q} + 2\zeta_{i,Q}$, where $L_{i,Q}$ and $\zeta_{i,Q}$ are the maximized likelihood and number of parameters for the ith whale, given Q. Thus, M_2 models the memory effects arbitrarily, since we have no a priori knowledge about these.

In an attempt to improve interpretability (see Table 1) while (greatly) reducing the number of parameters, M_1 is suggested,

$$M_1 : \operatorname{logit}(p_t) = \eta_t + \sum_{q=1}^{Q} c_\phi^*(q) y_{t-q}. \tag{3.10}$$

Here, the kernel $c_\phi^*(q)$ is a biexponential function,

$$c_\phi^*(q) = \phi_1 \exp\bigl(-\exp(\phi_2)q\bigr) + \phi_3 \exp\bigl(-\exp(\phi_4)q\bigr) \tag{3.11}$$

of $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)^T$. The factors $\exp(\phi_2)$ and $\exp(\phi_4)$ could theoretically be replaced by two positive constants; however, the form above allows for unconstrained optimization. In M_1, $\theta_1 = (\beta^T, \phi^T)^T$. The biexponential function is suggested because it provides a flexible class of *memory models* (models that include information about the past). Table 1 and Figure 3 summarize some possible shapes in dependence of the parameters; see also Li and Ditlevsen (2019) for simulated point processes under different kernels and external input. The kernel c_ϕ^* has the potential to capture the underlying trend of the autoregressive coefficients while stabilizing the autoregressive effects, forcing the memory component to be continuous and smooth. This will become apparent when presenting the results.

TABLE 1
Characteristics of the biexponential kernel for different parameter vectors ϕ in (3.11) used to model memory effects

Kernel	Description	Parameter	Interpretation
Bursting	first positive, then negative, then vanishing	$\phi_1 > 0 > \phi_3$, $\phi_1 > \lvert\phi_3\rvert$, $\phi_2 > \phi_4$	recent events have excitatory effects, accumulation of events has inhibitory effects, resulting in rhythmic bursts of events
Decaying	first negative, then vanishing	$\phi_1 = 0 > \phi_3$, $\lvert\phi_3\rvert$, $\exp(\phi_4)$ small	inhibitory effects are small but long-lasting, making event rate decay slowly over time
Delaying	first negative, then positive, then vanishing	$\phi_1 > 0 > \phi_3$, $\phi_1 < \lvert\phi_3\rvert$, $\phi_2 < \phi_4$	recent events have inhibitory effects, accumulation of events has excitatory effects, preventing short inter-event intervals
Exciting	first positive, then vanishing	$\phi_1, \phi_3 \geq 0$ $\max(\phi_1, \phi_3) > 0$	recent events have excitatory effects

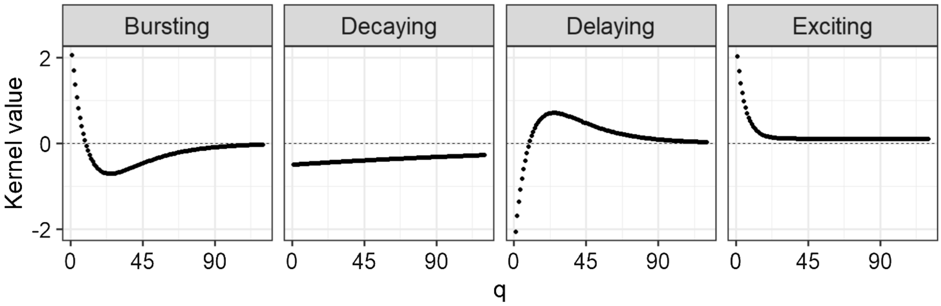

FIG. 3. *Characteristics of the biexponential kernel for different parameter vectors ϕ in (3.11) used to model memory effects.*

3.1. *Model selection.* For both buzzing and calling, a stepwise procedure based on pooled AIC is used to select the autoregressive order for M_2, starting at $Q = 0$ and then incrementing one at a time, for example, $Q = 1$, $Q = 2, \dots$ When the AIC has increased beyond 20, compared to the current minimum, the algorithm stops, and the Q resulting in the lowest AIC is chosen for M_2. This autoregressive order is then used when fitting M_1, and it is investigated if the model fit for M_1 results in a further AIC decrease.

3.2. *Diagnostics and performance.* For event times $u_1, \dots, u_J$, the inter-event intervals $u_{j+1} - u_j$ can be rescaled. They are denoted by $(z_j^i)_{j\in\{1,\dots,J-1\}, i\in\{0,1,2\}}$ and defined as

$$z_j^i = 1 - \exp\left(-\int_{u_j}^{u_{j+1}} \lambda_i\big(t \mid X(t), \hat{\theta}_i\big)\, dt\right), \tag{3.12}$$

with $j = 1, \dots, J-1$, $i = 0, 1, 2$, where $\hat{\theta}_i$ is the maximum likelihood estimator (MLE) of the parameter vector θ_i in model M_i. The rescaled z_j^i's will be independent and uniformly distributed (i.i.d.) random variables on the interval [0, 1) if and only if $\lambda_i(\cdot \mid \cdot, \hat{\theta}_i)$ equals the true data generating conditional intensity (Truccolo et al. (2005)). If a and b are the jth and

$(j+1)$th values of t where $Y_t = 1$, the integral in (3.12) is approximated as follows:

$$\begin{aligned}\int_a^b \lambda_i\big(t \mid X(t)\big)\,dt &\approx \sum_{t=a}^{b-1} \frac{\lambda_i(t \mid X(t)) + \lambda_i(t+1 \mid X(t+1))}{2}\Delta \\ &\approx \sum_{t=a}^{b-1} \frac{P_i(Y_t = 1 \mid X_t) + P_i(Y_{t+1} = 1 \mid X_{t+1})}{2}.\end{aligned} \tag{3.13}$$

Based on the z_j^i's, the Kolmogorov–Smirnov (KS) statistic can be used to test the significance of the disagreement between the hypothesized cumulative distribution function (CDF) F and the empirical CDF F_n which is given as $F_n(z) = \frac{1}{n}\sum_{j=1}^n 1_{(-\infty,z]}(Z_j)$ in stochastic form. For i.i.d. random variables $Z_1, \dots, Z_n$, the KS statistic is defined as $S_n = \sup_{z\in\mathbb{R}} |F_n(z) - F(z)|$. Under the hypothesis, $\sqrt{n}S_n$ converges in distribution to the Kolmogorov distribution. Large values are critical for the hypothesis. We have that $1 - P(\sqrt{n}S_n \leq 1.36) \approx 0.05$, and thus 1.36 is the critical value on significance level 0.05 (Wang, Tsang and Marsaglia (2003)).

KS confidence intervals are presented to help assess the concordance between model and data. The band defined by $F_n \pm 1.36/\sqrt{n}$ represents a uniform confidence band for the entire shape of F_n, meaning that F is entirely contained in the band if and only if the hypothesis cannot be rejected.

Potential violations of independence are investigated visually by plotting $z_2^i, z_3^i, \dots, z_{J-1}^i$ against $z_1^i, z_2^i, \dots, z_{J-2}^i$. If a nonrandom pattern can be observed, there is information propagating from one z_j^i to the next, and they are, therefore, not independent.

To check predictive performance in terms of the actual responses, a rolling validation scheme is implemented. When possible, a model is fit to approximately the first 50% of the data and then validated in the following 10%. Hereafter, the full 60% is used as training data and validated in the following 10%. This procedure is done five times, such that the last split has 90% as training data. For some whales the areas visited are not the same for all splits. In these cases the available information in the training set is taken into account, meaning that the area-variable is still used but could be a factor with fewer levels. If a new, unknown area is visited in the validation set, the algorithm will base its estimates on the last known area visited. For a given split the expected counts in each hour of the validation set is compared with the actual counts. Let B denote a bin corresponding to one hour, and let C_B denote the actual count during this hour. The expectation of the count is calculated as

$$E(C_B) = \sum_{t\in B} E(Y_t \mid X_t = x_t) = \sum_{t\in B} P(Y_t = 1 \mid X_t = x_t). \tag{3.14}$$

3.3. *Confidence intervals.* For $\theta, \psi \in \mathbb{R}^b$, the endpoints of an approximate 95% confidence interval for $\psi^T\theta$ are

$$\psi^T\hat{\theta} \pm 1.96 \cdot \sqrt{\psi^T\hat{\Sigma}(\hat{\theta})\psi}, \tag{3.15}$$

where $\hat{\Sigma}(\hat{\theta})$ is the estimated covariance matrix of $\hat{\theta}$. The logit at a specific setting of area, depth and time of day can be constructed by picking a suitable ψ.

3.4. *Implementation.* All computations were carried out using R version 3.6.1 (R Core Team (2019)). Unless stated otherwise, the default settings were utilized for the various functions.

In practice, the nonlinear functions g_D and g_H in (3.8) are approximated with splines. For g_D, the ns function from the splines library was used to estimate a natural cubic spline with three degrees of freedom, except for the whale Balder, where two degrees of freedom

TABLE 2
Starting values used for ϕ when fitting M_1

Sound	*Whale*	ϕ_1	ϕ_2	ϕ_3	ϕ_4
Buzzing	Balder	−7.53	−1.15	0.75	−3.76
	Eistla	−3.67	−1.52	0.25	−5.81
	Freya	−2.48	−1.93	0.19	−7.86
	Frida	−5.29	−2.26	2.80	−3.30
	Thora	−6.51	−0.96	0.76	−3.53
Calling	Balder	3.63	−2.00	−155.41	1.39
	Eistla	0.11	−14.18	2.33	−1.64
	Freya	16.57	−1.19	−43.77	−0.10
	Frida	8.22	−1.53	−23.44	−0.15
	Thora	4.76	−1.63	−10.52	−0.03

in the buzzing models were chosen to reach convergence. For g_H, the pbs function from the pbs library was chosen to estimate a periodic B-spline with three degrees of freedom and boundary knots at zero and 24 hours.

Both M_0 and M_2 were estimated utilizing glm, since they are generalized linear models (GLMs). To understand that M_2 is a also a GLM, recall that the posttagging silence and finite delay imply that there is a constant number of covariates for all observations; see (3.5). Because of the biexponential term, however, M_1 is not a GLM. Therefore, a general purpose optimizer (optim) was chosen to maximize the log-likelihood for this model (fnscale=-1 was utilized since optim performs minimization). The method used was BFGS, a quasi-Newton method. Specifically, the function to be maximized took θ_1 as input vector and calculated probabilities $p_1, \dots, p_T$ based on the inverse logit-function, after which it returned the joint probability according to (3.6). The starting vector for β when fitting M_1 was the MLE from M_2. Good choices of starting values for $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)^T$ were obtained with the help of the nls and SSbiexp functions. The values can be seen in Table 2.

When estimating M_1 during data splitting, the starting vector for β was the MLE from the M_2 fit to the training data, created by glm. Starting values for ϕ were those from the M_1 MLE based on the whole data. Additionally, maxit was set to 1000 when estimating in the calling model.

For the construction of confidence intervals, the vcov function was utilized to access the estimated covariance matrix if the MLE was obtained from glm. If instead it was obtained with the help of optim, the negation of the estimated hessian was inverted and utilized as the covariance matrix.

4. Results.

Figure 4 shows how the pooled AIC depends on Q when estimating in M_2. It turns out that, for buzzing, $Q = 68$ is the optimal choice, while, for calling, $Q = 41$ is best.

Using the best Q for M_2, it was investigated whether the AIC could be further lowered by instead fitting M_1. This was the case for buzzing, where a decrease of 275 was observed. For calling, however, this was not the case.

Model diagnostics using the z^i_j's are provided in Figures 5–6. KS confidence intervals are based on the asymptotic critical value, but bootstrapped critical values constructed by simulating 999 random variables from uniform distributions were similar. Judging by the KS-plots for buzzing in Figure 5, the distribution of $(z^i_j)_{j\in\{1,\dots,J-1\},i\in\{0,1,2\}}$ appears to become more uniform for all whales when introducing memory; this is seen for both M_1 and M_2. Here, the straight line is entirely contained in the bands for Eistla and Freya, and thus a formal hypothesis tests would not be able to reject that M_1 or M_2 is the true model for these

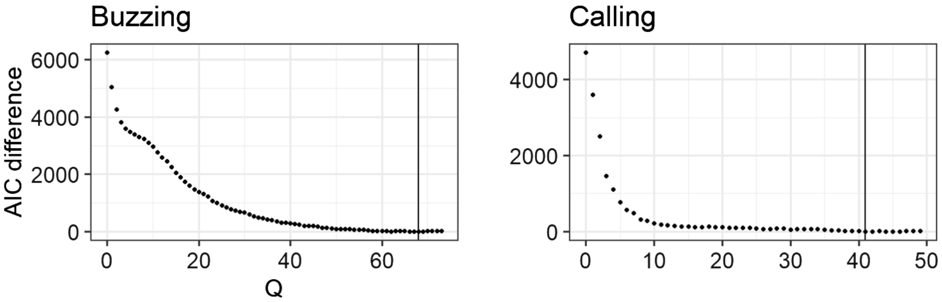

FIG. 4. *Pooled AIC increase as a function of the autoregressive order Q used when estimating in M_2; see (3.9). The vertical line indicate the Q for which the minimum AIC is attained.*

whales. Regarding Balder and Thora, the z_j^i's for the memory models have a distribution reasonably close to the uniform distribution; however, the large sample size exposes a small but significant difference. In the first three rows of Figure 6, one can assess the relationship between consecutive z_j^i's for buzzing. In general, the points seem to become more randomly scattered with memory, indicating a diminishing dependence and increased concordance.

The plots for calling in Figures 5–6 are less appealing than for buzzing, even though the distribution of $(z_j^i)_{j\in\{1,\dots,J-1\},i\in\{0,1,2\}}$ still become more uniform with memory. The wide confidence intervals in Figure 5 are a result of the low amount of calls. Despite the fact that a hypothesis test for the uniformity of the rescaled inter event times would be rejected for four out of five whales regardless of which memory model is used, it can still be seen in several plots that the confidence bands for M_2 are closer to containing the straight line than those for M_1 which is consistent with the fact that M_1 for calling was not a better memory model than M_2 according to the AIC. As can be observed in the final three rows of Figure 6, the fitted memory models for Eistla and Thora do not remove dependence between consecutive z's, indicating that the models are not entirely adequate.

Figure 7 shows how close the estimated expected counts per hour are to the actual counts, using the rolling validation scheme described in the Methods section. In the plots the validation sets are "glued" together. It should be mentioned that the Qs chosen for the entire data set (68 and 41) are fixed at these values for each split. Also considering Q as a parameter would make this pragmatic validation procedure very time consuming and error-prone. The estimates in Figure 7 are based on the best model fits, according to the AIC: M_1 for buzzing and M_2 for calling. The left column visualizes the performance of the buzzing models. Here, the majority of expected counts are reasonably close to the actual counts. There are some exceptions, for example, in the earlier part of Balder's plot where overestimation occurs. Later in the plot, where more data is used to learn from, it looks better. In general, the model's predictive performance does not seem poor. The sparsity of Frida's plot is because the fitting procedure did not work with 50%–80% training data. At 90%, however, convergence was attained. The right column in Figure 7 shows the expected counts and actual counts when calling is used as the response. Initially, it can be difficult to judge how M_2 fares since the call counts are so low during most hours. What can be said, however, is that the model appears to predict a very low call count when it actually is the case. Unfortunately, there are hours where the counts are far above what one would expect according to the model. These are not frequent but, nevertheless, indicate that there is room for improvement.

The MLEs will now be reported. The memory effects in M_1 and M_2 are presented together in Figure 8. The buzzing plots in the first row show a tendency for the coefficients in M_2 to start off negative, become slightly positive a little later on and then possibly decrease toward zero. This indicates that, among other things, a buzz is less likely to occur right after a previous buzz (refractory period), but there is a selfexcitatory effect at around 10–30 seconds

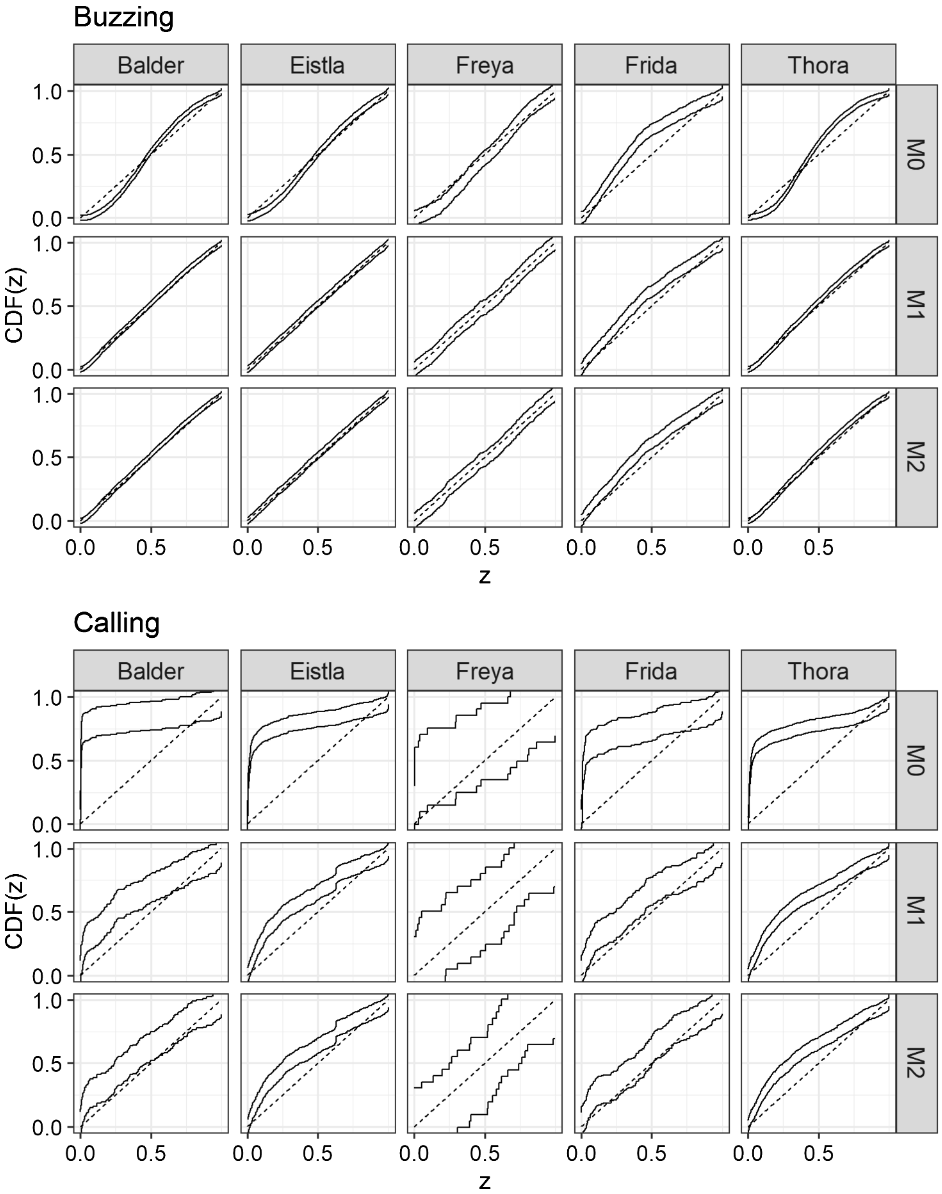

FIG. 5. *Model diagnostics for buzzing and calling using the rescaled interevent times* $z_1^i, \dots, z_{J-1}^i$, *as defined in* (3.12). *Kolmogorov–Smirnov confidence bands for the empirical CDF of the* z_j^i*'s are plotted together with the CDF of the uniform distribution. Diagnostics are shown for both model* M_0, M_1 *and* M_2; *see* (3.7)–(3.11).

after a buzz in agreement with a delaying kernel; see Table 1 and Figure 3. The biexponential sequence appears to capture the general trend of the coefficients well, even though it is obtained from the likelihood of M_1 and not estimated based on the coefficients.

The second row in Figure 8 visualizes the memory effects in the calling models. In Freya's plot the coefficients behave quite erratically. This can probably be attributed to the fact that there are only 21 calls in her record. Balder and Frida have more calls (146 and 149), and, while the coefficients still have a large variance, a possible underlying pattern begins to emerge. This pattern can again be spotted for Eistla and Thora, where the coefficients have a smaller variance, due to the larger number of calls (508 and 762). Apparently, the memory effect has a tendency to be positive early and then decrease toward zero, indicating that a call is more likely to occur a few seconds after a previous call. The curve in M_1 seems to model

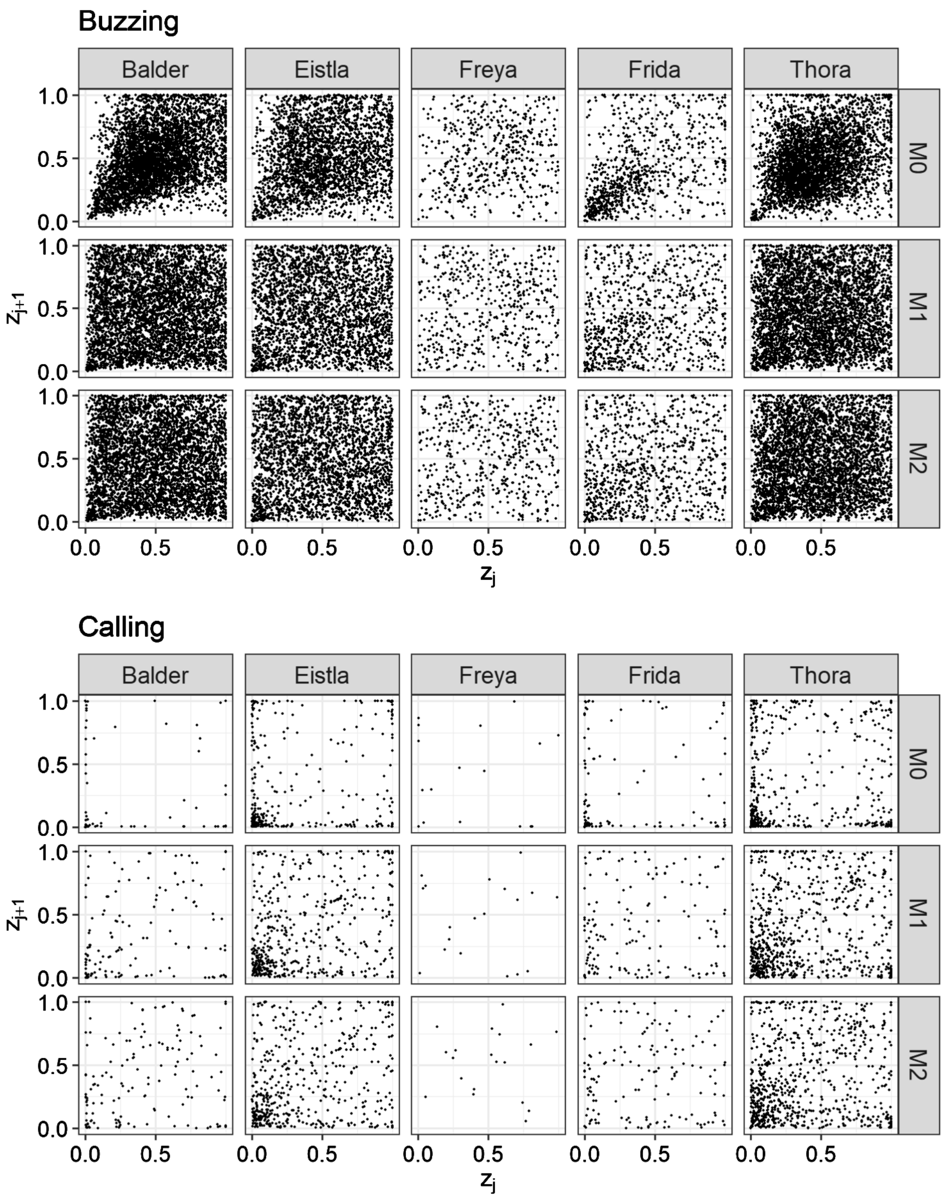

FIG. 6. *Model diagnostics for buzzing and calling using the rescaled interevent times* $z_1^i, \dots, z_{J-1}^i$, *as defined in* (3.12). *Here*, $z_2^i, \dots, z_{J-1}^i$ *are plotted against* $z_1^i, \dots, z_{J-2}^i$. *Diagnostics are shown for both model* M_0, M_1 *and* M_2; *see* (3.7)–(3.11).

the trend in an acceptable way; however, as mentioned, if one uses the AIC for model selection, M_1 is not advantageous compared to M_2 in terms of describing calling behaviour. For Eistla, $\hat{c}_\phi^*$ is an exciting kernel; see Table 1 and Figure 3. For Balder, $\hat{c}_\phi^*$ is already positive at a time lag of one second and thus, in practice, functions as an exciting kernel since the data is sampled at a 1 Hz rate. For Freya, Frida and Thora, the estimated kernel value is negative only at the first second.

In Figure 9, the effects of area, depth and time are illustrated for M_0 and the memory models of choice (M_1 for buzzing and M_2 for calling). Confidence intervals for the memory models are provided. They can be used as pragmatic measures of uncertainty but are only approximations, since they rely on the assumption that the corresponding suggested models are correct.

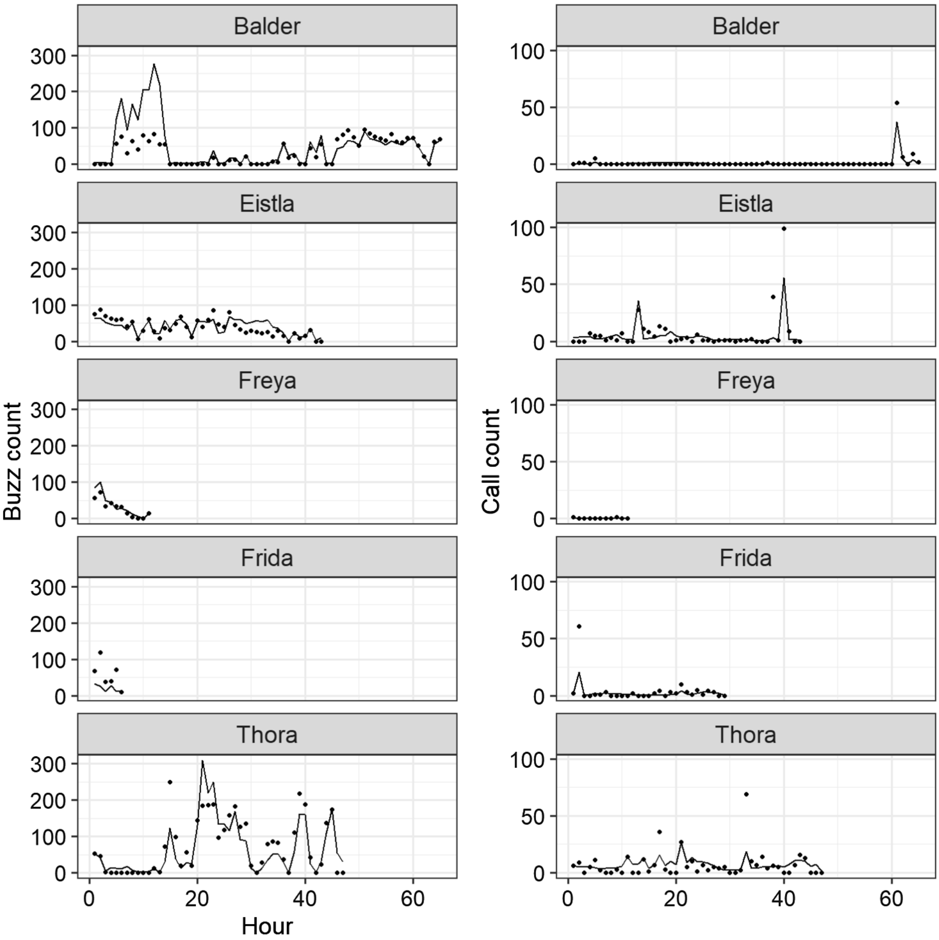

FIG. 7. *Actual counts (dots) and estimated expected counts (lines) based on M_1 for buzzing and M_2 for calling; see (3.8)–(3.11) and (3.14). All estimates are shown for left-out data.*

The first row in Figure 9 reports the estimated buzzing odds-ratios (ORs) with respect to reference areas (O for Balder and G for the others). All effects have the same directions in the models with and without memory (remain over/under one when introducing memory), and thus the conclusions regarding which area represents more/less buzzing do not change. Area G generally represents a high buzzing rate which could indicate a feeding area. Only for Thora is there an area (OG) that corresponds to a higher buzzing rate, and Balder is not observed in area G.

In rows 2–3, the effects of depth and time of day are visualized in area O for Balder and G for the other whales. In the depth plots the time is fixed at 15 (3 p.m.). In the time plots the depth is fixed at the common median for buzzing whales (361 m). Note that, since the models are additive, fixing the area and one of the continuous predictors will not change where the other continuous predictor causes the buzzing/calling intensity to decrease, increase, be minimized or maximized.

In the second row of Figure 9, the overall shape of the curve is preserved with memory, meaning that the depths where one expects the whale to buzz more/less are similar in the two models. At certain depths the curves overlap, and at others the magnitude of the effects are noticeably smaller in M_1. Therefore, the base model has a tendency to overestimate the effects of the extrinsic covariates compared to the memory model. In general, the buzzing intensity increases with increasing depth until at least 250 m for all whales. For Balder and

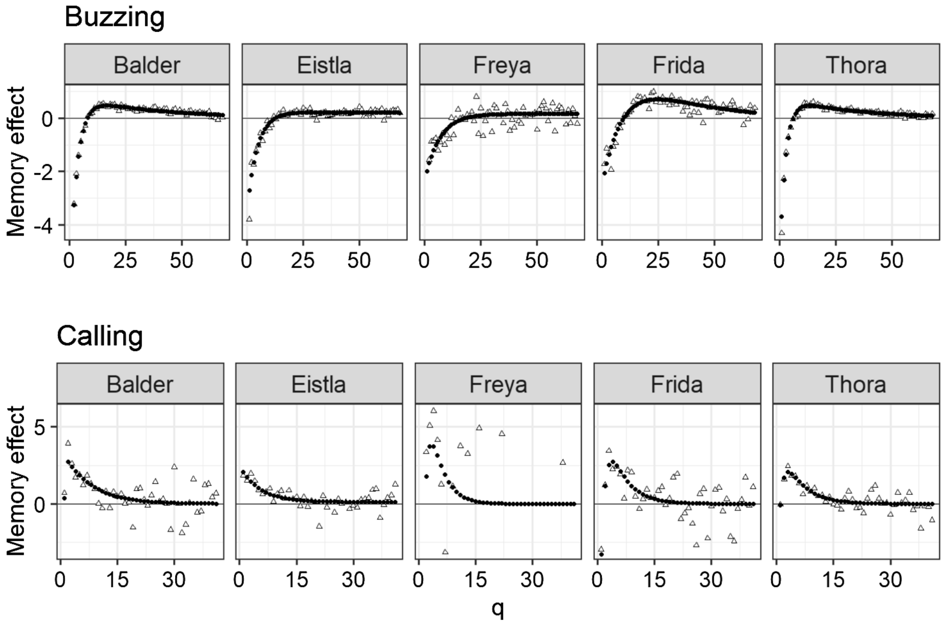

FIG. 8. *Autoregressive coefficients (white triangles) from model M_2 and the biexponential sequence (black dots) from model M_1; see (3.8)–(3.11). For visual reasons, some points with extreme values are not included.*

Frida the intensity peaks at a certain depth before falling again. It is also worth noticing that, even though area G generally represents a high buzzing probability, the highest buzzing probability in the depth plots for M_1 is attained by Balder in area O which is actually the area in which he buzzes the least. Thus, even though certain generalizations can be made regarding the depth effect, there is still considerable differences between whales.

The third row in Figure 9 can be used to assess the time-of-day effect on buzzing. Again, there are similarities in shape between the solid and dashed curve for a given whale, and overestimation occurs in M_0 for Balder, Frida and Thora. Generalizations of the time-of-day effect are difficult, as the whales behave differently. For example, Balder peaks in the evening, and Freya peaks in the morning while Frida and Thora peak around midnight.

The calling ORs are observable in the fourth row of Figure 9. It is noteworthy here that some ORs that before were under (over) 1 now are over (under) 1 in the memory models. This can be seen for Balder and Eistla when comparing area F to the reference area. Thus, the conclusions regarding which area represents more or less calls change for these whales when memory is added to the model.

Similarly to buzzing, including memory seems to muffle the calling intensity at certain depths, while at other depths the intensity is more or less the same, as can be seen in the fifth row of Figure 9. All models agree that there are more calls close to the surface; however, in a given area the calling intensity is also highly dependent on the whale. In area G, Thora calls much more than Eistla, who calls noticeably more than Freya and Frida. The broad confidence region for Freya is probably due to the very low number of calls in her data set (21 calls).

The final row of Figure 9 shows that it is also hard to generalize the time-of-day effect on calling. While there is some shape-preservation between M_0 and the memory models, how the intensity varies during the day is highly whale dependent.

5. Discussion. The results show that the logistic regression model presented in Blackwell et al. (2018) can be improved by adding a memory component.

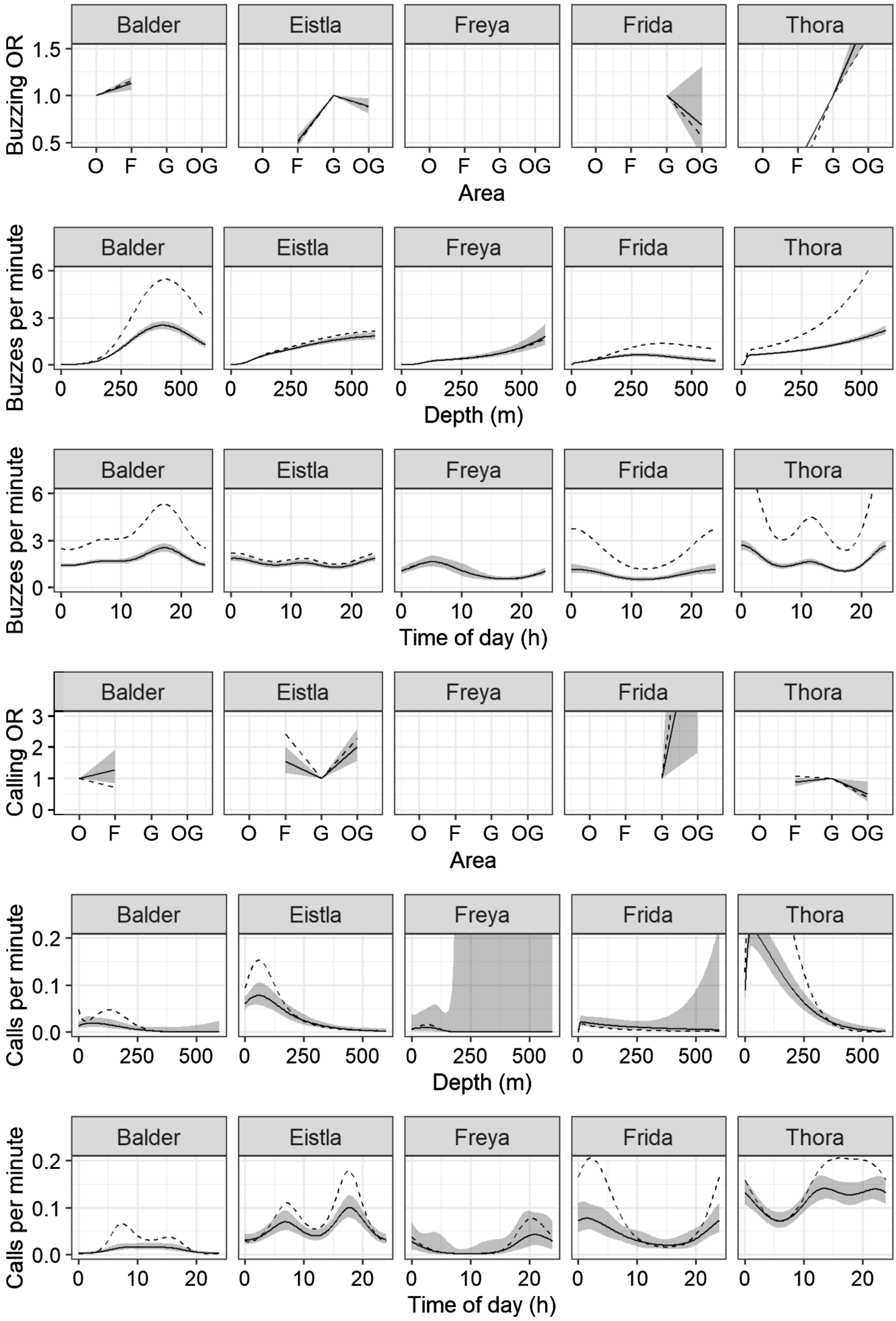

FIG. 9. *Odds-ratios and covariate effects. Rows* 1 *and* 4: *Odds-ratios for the areas. Rows* 2 *and* 5: *The depth effect. The time is fixed at* 15 (3 *p.m.*). *Rows* 3 *and* 6: *The effect of time. In row* 3 (6) *the depth is fixed at the common median for buzzing whales* (361 *m*) (*calling whales* (6.2 *m*)). *The dashed lines correspond to* M_0, *and the solid lines to* M_1 *for buzzing and* M_2 *for calling*; *see* (3.7)–(3.11). *Pointwise confidence intervals are provided for estimates in* M_1 *and* M_2.

For buzzing, using AIC for model selection, M_1 is the best model. Additionally, point process diagnostics reveal a considerably better concordance between model and data for M_1 than M_0, even though a significance based approach to the goodness-of-fit analysis concludes that M_1 is still not entirely satisfactory for all whales. A visual presentation of buzz count predictions indicates that M_1 has a decent predictive performance. There are some exceptions where the prediction do not match the observation well. This occasional unpredictability could perhaps be attributed to the fact that there are very small energetic costs of buzzing for narwhals (Noren et al. (2017)). Thus, a narwhal does not have to only buzz when it is strictly necessary, and some random deviation from more well-defined patterns is to be expected.

The memory model M_1 implies for all whales that, right after beginning buzzing, there is a lowered probability of beginning buzzing again. This can likely be explained by the fact that buzzes can last several seconds. Thus, the memory model has the capacity to implicitly take into account the buzz length, even if this is not included in the data set.

The memory model additionally implies that given a buzz at least seven to 19 seconds into the past (depending on the whale), there is a short period of slightly elevated buzzing probability. As such, even when conditioning on extrinsic covariates, M_1 allows the whale to enter a state of higher buzzing activity. As buzzing is used for feeding, such a buzzing mode could perhaps be triggered by encountering prey underwater. A possible objective for future analyses is to investigate whether there is an association between a given buzz series and successfully capturing a fish. Measurements of stomach temperature can be used as a predictor of the presence or absence of food in the stomach (Heide-Jørgensen et al. (2014)). Combining these measurements with Acousonde data could provide new information on the role of buzzing in feeding events.

Unlike with buzzing, AIC suggests that the memory model M_2 is the best model for calling. However, point process diagnostics were less appealing than with buzzing, even though they did improve for M_2, compared to M_0. The memory model indicates that, shortly after a previous call, there is an increased probability of observing another call. A possible explanation for this is that the calls in the data set can be from both the whale itself and other whales. As calls are presumably used for social communication, higher calling activity could, therefore, be the result of interaction between several whales in close proximity.

For both responses the effects of depth and time of day led to qualitatively similar conclusions with and without memory in terms of positive/negative contributions to buzzing/calling activity. The base model tends to overestimate the magnitude of the effects compared to the memory model. There is a considerable difference from whale to whale, but some similarities are apparent: The whales all prefer to buzz at deeper depths and call closer to the surface. Generalizations of the effects of time are harder to discern.

Regarding the area effects, buzzing activity is generally high in area G, indicating a possible feeding area. For calling, adding the autoregressive component changed the interpretation of certain areas in relation to expected calling intensity. While one should be careful with trusting the estimated ORs in the calling models, this qualitative discrepancy between M_0 and M_2 at least indicates that better fitting models and/or data sets with much higher numbers of calls are needed before clear area effects on calling behaviour can be claimed.

In general, it intuitively makes sense that there is some sort of dependency between observations for a given narwhal. The assumptions behind the base model are that of independent response variables given the extrinsic covariates which does not seem entirely reasonable; it does not take into account that the whales can buzz or call in certain patterns. However, it is important not to dismiss the base model based on its assumptions alone. As was already concluded, a lot of the extrinsic covariate effects are qualitatively quite similar in the models with and without memory. The base model is more easily interpretable and, possibly, a valuable addition to an exploratory analysis.

Regardless of whether or not a memory component is used, collecting extensive data sets from several whales and then fitting one model to the pooled data is a logical next step toward better understanding general narwhal behaviour. A random effects model could be useful here.

Acknowledgments. We thank Susanna Blackwell for supervising the analysis of narwhal buzzing activity.

REFERENCES

BLACKWELL, S., TERVO, O., CONRAD, A., SINDING, M.-H., HANSEN, R., DITLEVSEN, S. and HEIDE-JØRGENSEN, M. P. (2018). Spatial and temporal patterns of sound production in East Greenland narwhals. *PLoS ONE* **13** 06.

DERUITER, S. L., LANGROCK, R., SKIRBUTAS, T., GOLDBOGEN, J. A., CALAMBOKIDIS, J., FRIEDLAENDER, A. S. and SOUTHALL, B. L. (2017). A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *Ann. Appl. Stat.* **11** 362–392. MR3634328 https://doi.org/10.1214/16-AOAS1008

HEIDE-JØRGENSEN, M. P., DIETZ, R., LAIDRE, K. and RICHARD, P. (2002). Autumn movements, home ranges, and winter density of narwhals (Monodon monoceros) tagged in Tremblay Sound, Baffin Island. *Polar Biol.* **25** 05.

HEIDE-JØRGENSEN, M. P., NIELSEN, N., HANSEN, R. and BLACKWELL, S. (2014). Stomach temperature of narwhals (Monodon monoceros) during feeding events. *Anim. Biotelem.* **2** 06.

HEIDE-JØRGENSEN, M. P., NIELSEN, N., HANSEN, R., SCHMIDT, H. C., BLACKWELL, S. and JØRGENSEN, O. A. (2015). The predictable narwhal: Satellite tracking shows behavioural similarities between isolated subpopulations. *J. Zool.* **297** 09.

JOHNSON, M., AGUILAR DE SOTO, N. and MADSEN, P. T. (2009). Studying the behaviour and sensory ecology of marine mammals using acoustic recording tags: A review. *Mar. Ecol. Prog. Ser.* **395** 55.

KASS, R. E., EDEN, U. T. and BROWN, E. N. (2014). *Analysis of Neural Data. Springer Series in Statistics*. Springer, New York. MR3244261 https://doi.org/10.1007/978-1-4614-9602-1

KOBLITZ, J., STILZ, P., RASMUSSEN, M. and LAIDRE, K. L. (2016). Highly directional sonar beam of narwhals (Monodon monoceros) measured with a vertical 16 hydrophone array. *PLoS ONE* **11** 11.

LANGROCK, R., MARQUES, T. A., BAIRD, R. W. and THOMAS, L. (2014). Modeling the diving behavior of whales: A latent-variable approach with feedback and semi-Markovian components. *J. Agric. Biol. Environ. Stat.* **19** 82–100. MR3257903 https://doi.org/10.1007/s13253-013-0158-6

LI, K. and DITLEVSEN, S. (2019). Neural decoding with visual attention using sequential Monte Carlo for leaky integrate-and-fire neurons. *PLoS ONE* **14** 05.

LI, K., KOZYREV, V., KYLLINGSBÆK, S., TREUE, S., DITLEVSEN, S. and BUNDESEN, C. (2016). Neurons in primate visual cortex alternate between responses to multiple stimuli in their receptive field. *Front. Comput. Neurosci.* **10** 12.

NGO, M. C., HEIDE-JØRGENSEN, M. P. and DITLEVSEN, S. (2019). Understanding narwhal diving behaviour using hidden Markov models with dependent state distributions and long range dependence. *PLoS Comput. Biol.* **15**.

NOREN, D., HOLT, M., DUNKIN, R. and WILLIAMS, T. (2017). Echolocation is cheap for some mammals: Dolphins conserve oxygen while producing high-intensity clicks. *J. Exp. Mar. Biol. Ecol.* **495** 103.

PATTERSON, T. A., PARTON, A., LANGROCK, R., BLACKWELL, P. G., THOMAS, L. and KING, R. (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *AStA Adv. Stat. Anal.* **101** 399–438. MR3712406 https://doi.org/10.1007/s10182-017-0302-7

QUICK, N. J., ISOJUNNO, S., SADYKOVA, D., BOWERS, M., NOWACEK, D. P. and HIDDEN, A. J. R. (2017). Markov models reveal complexity in the diving behaviour of short-finned pilot whales. *Sci. Rep.* **7**.

RASMUSSEN, M., KOBLITZ, J. and LAIDRE, K. L. (2015). Buzzes and high-frequency clicks recorded from narwhals (Monodon monoceros) at their wintering ground. *Aquat. Mamm.* **41** 256.

TRUCCOLO, W., EDEN, U. T., FELLOWS, M., DONOGHUE, J. and BROWN, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* **93** 03.

WANG, J., TSANG, W. W. and MARSAGLIA, G. (2003). Evaluating Kolmogorov's distribution. *J. Stat. Softw.* **08** 12.

YUAN, Y., BACHL, F. E., LINDGREN, F., BORCHERS, D. L., ILLIAN, J. B., BUCKLAND, S. T., RUE, H. and GERRODETTE, T. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *Ann. Appl. Stat.* **11** 2270–2297. MR3743297 https://doi.org/10.1214/17-AOAS1078

R CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.